Der Wahlkampf 2019 in traditionellen und digitalen Medien: Technischer Bericht

Gilardi, Fabrizio ; Dermont, Clau ; Kubli, Mael ; Baumgartner, Lucien

DigDemLab

Der Wahlkampf 2019 in traditionellen und digitalen Medien

Code Buch & Technischer Report

Fabrizio Gilardi, Clau Dermont, Maël Kubli und Lucien Baumgartner

Digital Democracy Lab, Universität Zürich

Inhaltsverzeichnis

1	**Überblick Selects Medienanalyse Daten 2019**	**1**
1.1	Überblick	1
1.2	Datenquellen	1
1.2.1	Schweizerische Mediendatenbank (SMD)	1
1.2.2	Corriere del Ticino	2
1.2.3	Twitter	2
1.2.4	Facebook	3
2	**Datensätze**	**3**
2.1	Überblick Kandidierenden Listen	4
2.1.1	Ständeratsliste (2019_CHVOTE_STA)	4
2.1.2	Nationalratsliste (2019_CHVOTE_NAT)	5
2.2	Überblick SMD Daten	6
2.2.1	SMD Artikel klassifiziert mit Tonalität (smd_class_sent_MINI)	6
2.2.2	SMD Named Entity Datensatz (smd_ner_MINI_kand)	6
2.2.3	Named Entity Datensatz des CdT (cdt_ner_MINI_kand)	7
2.2.4	Named Entity Datensatz der Wahlprognosen (smd_ner_MINI_wahl)	8
2.3	Überblick Twitter Daten	9
2.3.1	Twitter Nachrichten (twitter_class_sent_MINI)	9
2.3.2	Twitter Hashtag der User (twitter_hashtags_MINI)	13
2.3.3	Twitter Statistiken (twitter_userstats_MINI)	15
2.3.4	Twitter wöchentliche Statistiken (twitter_w_userstats_MINI)	16
2.3.5	Twitter Friendslist (twitter_friendslist_MINI)	17
2.4	Überblick Facebook Daten	18
2.4.1	Facebook Pages (facebook_userstats_MINI)	18
2.4.2	Facebook Beiträge (facebook_class_sent_MINI)	19
3	**Technischer Bericht**	**19**
3.1	Kurzanleitung der automatisierten Medienanalyse	19
3.2	Framework	20
3.2.1	Datenbeschaffung	20
3.2.2	Tonalität	20
3.2.3	Klassifikation	20
3.2.4	Named Entity Recognition	20
3.2.5	Zusätzliche Schritte	20
3.3	Ingestion System	21
3.3.1	SMD (Schweizerische Mediendatenbank)	21
3.3.2	Twitter	22

	3.3.3	Facebook	23
3.4		Tonalität	23
3.5		Klassifikation	26
	3.5.1	Binärklassifikation (nur für SMD)	27
	3.5.2	Trainingsdaten	28
	3.5.3	Feature Engineering	30
	3.5.4	Ensemble Training	32
3.6		Named Entity Recognition (NER)	35
3.7		Netzwerkanalyse	36
4		**Referenzen**	**37**
5		**Appendix**	**39**
5.1		Liste der SMD Medien Quellen	39
5.2		Liste der Themen der Zeitungsartikel aus der SMD und des CdT	42
5.3		Liste der Themen zu den Beiträgen aus den sozialen Medien	43

1 Überblick Selects Medienanalyse Daten 2019

1.1 Überblick

Für die Selects Medienanalyse 2019 (Gilardi et al. 2020) wurden neben den Daten der schweizerischen Medien Datenbank neu auch Daten von Sozialen Medien analysiert. Dieser Teil umfasst Daten von Facebook und Twitter. Die Erhebungsperiode aller Daten umfasst den Zeitraum vom 01.01.2019 bis 31.10.2019. Das Ziel der Medienanalyse war es die wichtigsten AkteurInnen (Personen und Parteien) während des Wahlkampfes zu identifizieren, sowie zu untersuchen welche Themen während des Wahlkampfes in den Medien gehör fanden.

1.2 Datenquellen

Die Selects Medienanalyse erfordert das Analysieren von grossen Mengen textbasierter Daten über eine lange Zeitperiode. Dieser Umstand veranlasste uns, die Daten zentralisiert und vollautomatisiert zu erheben, da dies ein erheblicher Zeitgewinn gegenüber dem manuellen Erstellen der Datensätze darstellt.
Das DigDemLab besitzt ein stabiles Framework zur automatischen Verarbeitung und Sicherung für eine grosse Anzahl Daten von unterschiedlichster Art in Echtzeit, sowie die Kompetenzen Daten von neuen Onlinequellen zu beziehen und zu verarbeiten. Deshalb konnten wir nicht nur, wie bei vorherigen Selects Medienanalysen, die klassischen Medien bestehend aus Zeitungen analysieren, sondern auch die Wahlkampfperiode auf den sozialen Medien verfolgen.

1.2.1 Schweizerische Mediendatenbank (SMD)

Im Rahmen der Selects Medienanalyse erhielten wir Zugriff auf die Artikel von 86 verschiedenen Zeitung die all ihre Artikel in der schweizerischen Mediendatenbank (SMD) hinterlegen. Diese 86 Zeitungen publizierten im Zeitfenster der Analyse 1’141’053 Artikel. Der Korpus setzt sich aus insgesamt 891’996 (78.2 %) Artikeln auf Deutsch, 236’731 (20.8 %) Artikeln auf Französisch, 11’349 (1 %) Artikeln auf Italienisch, 947 Artikeln auf Englisch und 30 Artikeln auf Rätoromanisch zusammen. All diese Artikel wurden in mehreren teilautomatisierten und vollautomatisierten Schritten klassifiziert,

analysiert und ausgewertet. Die Artikel der Zeitungen wurden während des Erhebungszeitraumes täglich aktualisiert, wobei die neuen Artikel direkt in einer Datenbank abgespeichert, verarbeitet und vorbereitet wurden für die Auswertungen der verschiedenen Analysen. Je nach Auswertungsstrategie wurden verschiedene Sprachen berücksichtigt, wobei die Englischen und Rätoromanischen Artikel aufgrund der kleinen Anzahl nicht berücksichtigt wurden.

1.2.2 Corriere del Ticino

Zusätzlich zu den Publikationen der SMD haben wir noch alle Zeitungsartikel vom Corriere del Ticino gesammelt, um wenigsten eine italienischsprachige Tageszeitung analysieren zu können. Dieser Datensatz umfasst 6'628 Artikel über den Erhebungszeitraum.

1.2.3 Twitter

Für die Analyse des Wahlkampfes auf Twitter sammelten wir alle Tweets von Kandidierenden, Parteien, Departemente und ausgewählten Verbänden und Zeitungen über Twitters gut zugängliche API (Application Programming Interface). Wie schon bei der vorherigen Datenquelle wurden die Daten vom 01.01.2019 - 31.10.2019 erhoben. Dies umfasst 1'284 Accounts, die ein öffentliches Konto auf Twitter hatten. Hierbei handelt es sich um 1'239 Accounts von Kandidierenden, was ca. 27 % aller Kandidierenden entspricht, sowie 20 Parteien, 18 Zeitungen, 7 Organisationen und den 7 Departementen, sowie die Accounts von den Parlamentsdiensten. All diese Accounts zusammen setzten während dieses Zeitraumes 249'818 Tweets ab. Diese wurden, wie schon die Daten der Zeitungsartikel, täglich aktualisiert und prozessiert. Darüber hinaus sammelten wir während des Verlaufs des Projektes weitere 110'532 Tweets, die gewisse Schlüsselwörter (Hashtags) enthielten, die im Zusammenhang mit den Nationalen Wahlen und Abstimmungen im Jahr 2019 standen. Insgesamt umfasst unser Datensatz von Twitter damit 360'341 Tweets.

1.2.4 Facebook

Im Gegensatz zu Twitter stellte sich die Datenerhebung auf Facebook als viel komplexer heraus, da Facebook keine API mehr anbietet, die es einem erlaubt Beiträge von verschiedenen Accounts zu sammeln. Daher mussten wir auf eine externe Quelle zurückgreifen, die einen Zugang zu Facebooks Enterprise API besitzt.

Dieser Zugang wurde uns schlussendlich von der Socialmedia Management Firma FanpageKarma (fanpagekarma.com) gewährt. Die Plattform ermöglichte es uns zumindest alle Pages von Kandidierenden zu überwachen. Facebook unterscheidet bei Accounts zwischen privaten Sites und öffentlichen Pages. Dabei stellte sich heraus, dass selbst für Firmen, wie FanpageKarma nur Daten von öffentlichen Pages abgegriffen werden können, weshalb wir nur von 261 Kandidierenden Daten sammeln konnten. Die anderen Kandidierenden haben nur eine private Seite eingerichtet (Ende September 2019). Die Kennzahlen der einzelnen Pages und deren Beiträge, die von FanpageKarma zur Verfügung gestellt werden, haben wir am Ende mit einem teilautomatisierten Webscraper heruntergeladen und die Daten gleich für unsere Verwendungszwecke vorbereitet. Insgesamt umfasst dieser Datensatz 20’893 Beiträge von Facebook vom 01.01.2019 - 31.10.2019.

2 Datensätze

Die unten aufgeführte Tabelle 1 gibt einen Überblick über die 13 verschiedenen Datensätze. Jeder Datensatz ist für eine oder mehrere Arten von Analyse zu verwenden. Die Datensätze können zum Teil über gemeinsame Variablen miteinander verknüpft werden, sofern sie von derselben Quelle stammen. Dies kann über Dokumentennummern geschehen bei Daten aus den sozialen Medien oder über andere gemeinsame Variablen in den Datensätzen deren Grundlage die Zeitungsartikel von der SMD sind.

Datensatz	Formate	Inhalt
2019_CHVOTE_STA	.csv	Liste aller Kandidierenden Ständerat
2019_CHVOTE_NAT	.csv	Liste aller Kandidierenden Nationalrat
SMD_CLASS_SENT_MINI	.csv/.RDS	Alle Zeitungsartikel mit Sentiment und Klassifikation
SMD_NER_MINI_KAND	.csv/.RDS	Alle Nennungen von PolitikerInnen im SMD Datensatz
CDT_NER_MINI_KAND	.csv/.RDS	Alle Nennungen von PolitikerInnen im Datensatz des Correire del Ticino
SMD_NER_MINI_WAHL	.csv/.RDS	Alle Nennungen von Wahlprognosen im SMD Datensatz
TWITTER_CLASS_SENT_MINI	.csv/.RDS	Alle Tweets mit Sentiment und Klassifikation
TWITTER_HASHTAGS_MINI	.csv/.RDS	Meistbenutzte Hashtags der Kandidierenden
TWITTER_USERSTATS_MINI	.csv/.RDS	Twitter Metadaten der Kandidierenden
TWITTER_W_USERSTATS_MINI	.csv/.RDS	Statistische Kennzahlen Twitter UserInnen pro Woche
TWITTER_FRIENDSLIST_MINI	.csv/.RDS	Freundeslisten der Kandidierenden für die Netzwerkanalyse
FACEBOOK_USERSTATS_MINI	.csv/.RDS	Alle Metadaten der Facebook Pages
FACEBOOK_CLASS_SENT_MINI	.csv/.RDS	Alle Facebook Beiträge mit Sentiment und Klassifikation

Tabelle 1:Überblick über die wichtigsten Datensätze

2.1 Überblick Kandidierenden Listen

Insgesamt umfassen die beiden Datensätze 4'660 Kandidatinnen und Kandidaten, wovon 162 Ständeratskandidierende sind und 4'498 Nationalratskandidierende sind. Bei den Ständeratskandidierenden sind 35 % der Personen Frauen und 65 % der Personen Männer, bei den Nationalratskandidierenden sind es 40 % und 60 %.

2.1.1 Ständeratsliste (2019_CHVOTE_STA)

Variable Name	Beschreibung
firstname	Vorname des Kandidaten/ der Kandidatin
lastname	Nachname des Kandidaten/ der Kandidatin
gender	Geschlecht des Kandidaten/ der Kandidatin
year_of_birth	Geburtsjahr des Kandidaten/ der Kandidatin
age	Alter in Jahren des Kandidaten/ der Kandidatin
zip	Postleitzahl des Wohnortes des Kandidaten/ der Kandidatin
city	Wohngemeinde des Kandidaten/ der Kandidatin
country	Angabe des Wohnsitz-Land (Variable zur Identifikation von Auslandschweizern)

party_short	Partei Name in Kurzform des Kandidaten/ der Kandidatin
canton	Wohnkanton (Heimatkanton) des Kandidaten/ der Kandidatin
incumbent	Bisherigen Status des Kandidaten / der Kandidatin
link_personal_website	Webadresse der Persönlichen Webseite des Kandidaten / der Kandidatin
link_facebook	Link zur Facebook Page oder Site des Kandidaten / der Kandidatin
link_twitter	Link zum Twitter Account des Kandidaten / der Kandidatin
link_instagram	Link zum Instagram Account des Kandidaten / der Kandidatin

Tabelle 2: Variablen Beschreibung des Ständeratskandidatendatensatzes

2.1.2 Nationalratsliste (2019_CHVOTE_NAT)

Variable Name	Beschreibung
firstname	Vorname des Kandidaten/ der Kandidatin
lastname	Nachname des Kandidaten/ der Kandidatin
gender	Geschlecht des Kandidaten/ der Kandidatin
year_of_birth	Geburtsjahr des Kandidaten/ der Kandidatin
age	Alter in Jahren des Kandidaten/ der Kandidatin
zip	Postleitzahl des Wohnortes des Kandidaten/ der Kandidatin
city	Wohngemeinde des Kandidaten/ der Kandidatin
country	Angabe des Wohnsitz-Land (Variable zur Identifikation von Auslandschweizern)
party_short	Partei Name in Kurzform des Kandidaten/ der Kandidatin
canton	Wohnkanton (Heimatkanton) des Kandidaten/ der Kandidatin
list	Listenname auf der der Kandidat / die Kandidatin aufgeführt ist
list_place_1	Erster Listenplatz des Kandidaten / der Kandidatin
list_place_2	Zweiter Listenplatz des Kandidaten / der Kandidaten sofern vorhanden
incumbent	Bisherigen Status des Kandidaten / der Kandidatin
link_personal_website	Webadresse der Persönlichen Webseite des Kandidaten / der Kandidatin
link_facebook	Link zur Facebook Page oder Site des Kandidaten / der Kandidatin
link_twitter	Link zum Twitter Account des Kandidaten / der Kandidatin
link_instagram	Link zum Instagram Account des Kandidaten / der Kandidatin

Tabelle 3: Variablen Beschreibung des Nationalratskandidatendatensatzes

2.2 Überblick SMD Daten

2.2.1 SMD Artikel klassifiziert mit Tonalität (smd_class_sent_MINI)

Variable Name	Beschreibung
so	Zeitungskürzel
so_txt	Zeitungsname
pubDateTime	Publikationsdatum des Artikels
la	Sprache des Artikels
ru	Rubrik des Artikels falls vorhanden
ht	Titel des Artikels
ut	Untertitel des Artikels / Lead des Artikels
url	URL des Artikels falls vorhanden (nur bei Online Zeitungen)
annotation_geography	Angabe des oder der Länder von dem der Artikel handelt
annotation_person	Angabe der Personen die im Artikel vorkommen, sofern vorhanden
selectsclass	Themen Klassifikation des Artikels
sentiment_value	Tonalitätswert des Artikels

Tabelle 4: Variablen Beschreibung des Zeitungsartikeldatensatzes von der SMD

2.2.2 SMD Named Entity Datensatz (smd_ner_MINI_kand)

Variable Name	Beschreibung
doc_id	Dokumentennummer vom NER Prozess
person_id	Personennummer vom NER Prozess
so	Zeitungskürzel
so_txt	Zeitungsname
pubDateTime	Publikationsdatum des Artikels
la	Sprache des Artikels
annotation_geography	Angabe des oder der Länder von dem der Artikel handelt
annotation_person	Angabe der Personen die im Artikel vorkommen, sofern vorhanden
selectsclass	Themen Klassifikation des Artikels
sentiment_value	Tonalitätswert des Artikels
firstname	Vorname des genannten Kandidaten / Kandidatin
lastname	Nachname des genannten Kandidaten / Kandidatin
gender	Geschlecht des genannten Kandidaten / Kandidatin
year_of_birth	Geburtsjahr des genannten Kandidaten / Kandidatin falls vorhanden
age	Alter des genannten Kandidaten / Kandidatin falls vorhanden
zip	Postleitzahl des Wohnortes des genannten Kandidaten / Kandidatin

city	Wohngemeinde des genannten Kandidaten / Kandidatin
language	Muttersprache des genannten Kandidaten / Kandidatin
party_short	Partei des genannten Kandidaten / Kandidatin
canton	Wohnkanton des genannten Kandidaten / Kandidatin
list	Wahllistenname auf dem der genannte Kandidat / Kandidatin aufgeführt ist
list_place	Wahllistenplatz auf dem der genannte Kandidat / Kandidatin aufgeführt ist
incumbent	Handelt es sich bei der genannten Person im Artikel um ein Bisheriges Ratsmitglied
region	Gruppierung der Kantone in fünf verschiedene Grossregionen, Zürich und das Tessin
link_Twitter	Link zur Twitterseite des genannten Kandidaten / Kandidatin
link_facebook	Link zur Facebookseite des genannten Kandidaten / Kandidatin
link_Instagram	Link zur Instagramseite des genannten Kandidaten / Kandidatin
link_personal_website	Link zur Webseite des genannten Kandidaten / Kandidatin
candidacy	Für welche Kammer oder Kammern (Ständerat und Nationalrat) der genannt Kandidat / Kandidat zur Wahl antritt
bundesrat	Handelt es sich bei der im Artikel genannten Berson um einen Bundesrat oder nicht
candidate	Tritt die Person zur Wahl an
fullname	Vorname und Nachname der genannten Person
name	Regex Suchbegriff des Namens der genannten Person im Artikel

Tabelle 5: Variablen Beschreibung des Datensatzes der Nennungen der Kandidaten in den Zeitungsartikeln von der SMD

2.2.3 Named Entity Datensatz des CdT (cdt_ner_MINI_kand)

Variable Name	Beschreibung
doc_id	Dokumentennummer vom NER Prozess
person_id	Personennummer vom NER Prozess (identisch mit Nummer vom SMD NER)
so	Zeitungskürzel
so_txt	Zeitungsname
pubDateTime	Publikationsdatum des Artikels
la	Sprache des Artikels
sentiment_value	Tonalitätswert des Artikels
firstname	Vorname des genannten Kandidaten / Kandidatin
lastname	Nachname des genannten Kandidaten / Kandidatin
gender	Geschlecht des genannten Kandidaten / Kandidatin
year_of_birth	Geburtsjahr des genannten Kandidaten / Kandidatin falls vorhanden
age	Alter des genannten Kandidaten / Kandidatin falls vorhanden
zip	Postleitzahl des Wohnortes des genannten Kandidaten / Kandidatin

city	Wohngemeinde des genannten Kandidaten / Kandidatin
language	Muttersprache des genannten Kandidaten / Kandidatin
party_short	Partei des genannten Kandidaten / Kandidatin
canton	Wohnkanton des genannten Kandidaten / Kandidatin
list	Wahllistenname auf dem der genannte Kandidat / Kandidatin aufgeführt ist
list_place	Wahllistenplatz auf dem der genannte Kandidat / Kandidatin aufgeführt ist
incumbent	Handelt es sich bei der genannten Person im Artikel um ein Bisheriges Ratsmitglied
region	Gruppierung der Kantone in fünf verschiedene Grossregionen, Zürich und das Tessin
link_Twitter	Link zur Twitterseite des genannten Kandidaten / Kandidatin
link_facebook	Link zur Facebookseite des genannten Kandidaten / Kandidatin
link_Instagram	Link zur Instagramseite des genannten Kandidaten / Kandidatin
link_personal_website	Link zur Webseite des genannten Kandidaten / Kandidatin
candidacy	Für welche Kammer oder Kammern (Ständerat und Nationalrat) der genannt Kandidat / Kandidat zur Wahl antritt
bundesrat	Handelt es sich bei der im Artikel genannten Berson um einen Bundesrat oder nicht
candidate	Tritt die Person zur Wahl an
fullname	Vorname und Nachname der genannten Person
name	Regex Suchbegriff des Namesn der genannten Person im Artikel

Tabelle 6: Variablen Beschreibung des Datensatzes der Nennungen der Kandidaten in den Zeitungsartikeln des Correire del Ticino

2.2.4 Named Entity Datensatz der Wahlprognosen (smd_ner_MINI_wahl)

Variable Name	Beschreibung
doc_id	Dokumentennummer vom NER Prozess
txt	Dokumenten Inhalt (Text)
match_id	Match Identifikationsnummer
match_regex	Name des gefundenen Wahlprognoseinstitutes oder der gefundenen Person die Wahlprognosen abliefert.
so	Zeitungskürzel
so_txt	Zeitungsname
pubDateTime	Publikationsdatum des Artikels
la	Sprache des Artikels
ru	Rubrik des Artikels falls vorhanden
ht	Titel des Artikels
ut	Untertitel des Artikels / Lead des Artikels
url	URL des Artikels falls vorhanden (nur bei Online Zeitungen)
annotation_geography	Angabe des oder der Länder von dem der Artikel handelt

annotation_person	Angabe der Personen die im Artikel vorkommen, sofern vorhanden
selectsclass	Themen Klassifikation des Artikels
sentiment_value	Tonalitätswert des Artikels

Tabelle 7: Liste der Nennungen alle Institute die Wahlbefragungen und dergleichen im Jahr 2019 durchgeführt haben, sowie alle Experten dieser Institute, die für Zeitungen Wahlprognosen abgeliefert haben.

2.3 Überblick Twitter Daten

2.3.1 Twitter Nachrichten (twitter_class_sent_MINI)

Variable Name	Beschreibung
Akteur.Typ	Typ des Accounts (Person, Party, Organisation, usw.)
Akteur	Name der Organisation
Kurzel	Abkürzung des Organisationsnamens
First_Name	Vorname der Person
Last_Name	Nachname der Person
Gender	Geschlecht der Person
Year.of.birth	Geburtsjahr der Person
Age	Alter der Person in Jahren
Language	Sprache die die Person spricht
Canton	Wohnkanton
District	Bezirksname in dem die Person ihren Wohnsitz hat
Municipality	Wohnort der Person
Zip	Postleitzahl des Wohnortes der Person
National	Aktiv in Nationaler Politik
Regional	Aktiv in Regionale Politik
Chamber	Ratskammer in der Person Sitz inne hat
Incumbent	Bisherigen Status einer Person (Kandidaten)
Candidate.Kantonsrat	Binäre Variable, die angibt ob Person für ein Kantonstrat kandidiert
Candidate.Regierungsrat	Binäre Variable, die angibt ob Person für ein Regierungsrat kandidiert
Candidate.Nationalrat	Binäre Variable, die angibt ob Person für ein Nationalrat kandidiert
Candidate.Ständerat	Binäre Variable, die angibt ob Person für ein Ständerat kandidiert
Fraction	Fraktionszugehörigkeit im Bundeshaus, sofern es sich um eine Person handelt, die einen Sitz im Parlament innehält.
Party	Name der Mutterpartei der die Person angehört
Party_Short	Kürzel der Partei, der die Person angehört (Detailliert)
Quelle	Quelle der Personendaten

Selectsclass	Themen Klassifikation des Beitrages auf Twitter
Sentiment_value	Tonalitätswert des Beitrages auf Twitter
positive_words	Vektor der positiven Schlüsselwörter im Beitrag
negative_words	Vektor der negativen Schlüsselwörter im Beitrag
Datum	Publikationsdatum des Beitrages ohne Uhrzeit
Datum_full	Publikationsdatum des Beitrages mit Uhrzeit
Account_created_at	Erstellungsdatum des Accounts auf Twitter
Account_lang	Nutzerspracheinstellung auf Twitter
Country	Landeseinstellung des Accounts, falls Lokalisierung vom Nutzer erlaubt
Country_code	Landesabkürzung des Accounts, falls Lokalisierung vom Nutzer erlaubt
Description	Die benutzerdefinierte UTF-8-Zeichenkette, die das Konto beschreibt (Annullierbar)
Ext_media_expanded_url	Vollständige URL von externen Medienquellen
Ext_media_type	Typ der externen Medienquellen
Ext_media_url	Kurze URL von externen Medienquellen
Favorite_count	Zeigt ungefähr an, wie oft dieser Tweet von Twitter-Nutzern gemocht wurde
Favourites_count	Die Anzahl der Tweets, die dieser Benutzer im Laufe der Lebensdauer des Kontos gemocht hat.
Followers_count	Die Anzahl der Anhänger, die dieses Konto derzeit hat
Friends_count	Die Anzahl der Benutzer, denen dieses Konto folgt (AKA ihre «Gefolgschaft»).
Hashtags	Vektor aller Hashtags die im Tweet vorkommen
Is_quote	Ist der Beitrag ein Zitat eines anderen Beitrages (Quote)
Is_retweet	Ist der Beitrag ein Retweet eines anderen Beitrags (Retweet)
La	Sprache des Beitrages (Automatisiertes Verfahren von Twitter)
Lang	Spracheinstellung des Accounts
Listed_count	Die Anzahl der öffentlichen Listen, in denen dieser Benutzer Mitglied ist
Location	Der benutzerdefinierte Standort für das Profil dieses Kontos. Dieses Feld wird vom Suchdienst gelegentlich unscharf interpretiert.
Media_expanded_url	Vollständige URL von im Beitrag eingebetteten Medien
Media_type	Medientyp des eingebundenen URLs
Media_url	Kurze URL vom in Beitrag eingebetteten Medien
Mentions_screen_name	Vektor, der alle Anzeigenamen der Nutzer enthält, die im Beitrag erwähnt werden.
Mentions_user_id	Vektor, der alle Nutzer-IDs der Nutzer enthält, die im Beitrag erwähnt werden.
Name	Benutzername des Accounts (Statisch)

Place_full_name	Falls Lokalisierung vom Nutzer zugelassen ist, zeigt die den Ortsnamen mit Region an
Place_name	Falls Lokalisierung vom Nutzer zugelassen ist, zeigt dies den Ortsnamen an
Place_type	Falls Lokalisierung vom Nutzer zugelassen ist, zeigt dies die Art des Ortes an
Place_url	Falls Lokalisierung vom Nutzer zugelassen ist, zeigt dies die URL mit den zusätzlichen Metadaten des Ortes an
Profile_background_url	URL des Hintergrundes des Accounts
Porfile_banner_url	URL des Profilbanners des Accounts
Profile_extpanded_url	Vollständiges URL des Accounts
Profile_image_url	URL des Profilbildes des Accounts
Profile_url	Kurzes URL des Accounts
Protected	Binäre Variable die Anzeigt, ob der Account Privat ist (öffentlich zugänglich)
Screen_name	Anzeigename des Accounts (nicht Statisch)
Source	Dienstprogramm, mit dem der Tweet als HTML-formatierter String gepostet wird.
Status_id	Einzigartige Identifikationsnummer eines Beitrages (Tweets)
Status_url	URL des Beitrages
Statuses_count	Anzahl Beiträge die der Account abgesetzt hat.
Symbols	Spezielle Symbole und Emojis die im Text des Tweets vorkommen
Text	Text des Beitrages
Url	Eine URL, die vom Benutzer in Verbindung mit seinem Profil angegeben wird
Urls_expanded_url	-
Urls_url	-
User_id	Die ganzzahlige Darstellung des eindeutigen Identifikators für diesen Tweet. Diese Zahl ist größer als 53 Bit und einige Programmiersprachen können Schwierigkeiten/Störungen bei der Interpretation haben. Die Verwendung einer vorzeichenbehafteten 64-Bit-Ganzzahl zur Speicherung dieses Identifiers ist sicher, wie auch die Speicherung als String.
Verified	Wenn der Wert «TRUE» ist, wird angezeigt, dass der Benutzer ein verifiziertes Konto hat.
Reply_to_screen_name	Wenn der dargestellte Tweet eine Antwort ist, enthält dieses Feld den Bildschirmnamen des Autors des ursprünglichen Tweets.
Reply_to_status_id	Wenn der dargestellte Tweet eine Antwort ist, enthält dieses Feld die ganzzahlige Darstellung der ID des ursprünglichen Tweets.

Reply_to_user_id	Wenn der dargestellte Tweet eine Antwort ist, enthält dieses Feld die ganzzahlige Darstellung der Autoren-ID des ursprünglichen Tweets. Dies wird nicht unbedingt immer der im Tweet direkt erwähnte Benutzer sein.
Quoted_created_at	Publikationsdatum des Zitierten Beitrages mit Uhrzeit
Quoted_description	Die benutzerdefinierte UTF-8-Zeichenkette, die das Konto beschreibt (Annullierbar) von dem das Zitat stammt
Quoted_favorite_count	Zeigt ungefähr an, wie oft dieser Tweet von Twitter-Nutzern gemocht wurde
Quoted_followers_count	Die Anzahl der Anhänger die das Konto, von dem das Zitat stammt, derzeit hat
Quoted_friends_count	Die Anzahl der Benutzer, denen das Konto, von dem das Zitat stammt, folgt (AKA ihre «Gefolgschaft»).
Quoted_location	Der benutzerdefinierte Standort für das Profil dieses Kontos. Dieses Feld wird vom Suchdienst gelegentlich unscharf interpretiert.
Quoted_name	Benutzername des Accounts (Statisch)
Quoted_retweet_count	Wie oft der Beitrag, von dem das Zitat stammt, schon retweetet wurde.
Quoted_screen_name	Anzeigename des Accounts (nicht Statisch)
Quoted_source	Dienstprogramm, mit dem der Tweet als HTML-formatierter String gepostet wird.
Quoted_status_id	Dieses Feld erscheint nur, wenn der Tweet ein Zitat-Tweet ist. Dieses Feld enthält den ganzzahligen Wert Tweet ID des zitierten Tweets.
Quoted_statuses_count	Anzahl Beiträge die der Account, von dem das Zitat stammt, abgesetzt hat.
Quoted_text	Dieses Feld erscheint nur, wenn der Tweet ein Zitat-Tweet ist. Dieses Attribut enthält das Tweet-Objekt des Original-Tweets, das in einem Zitat steht.
Quoted_user_id	Die ganzzahlige Darstellung des eindeutigen Identifikators für diesen Tweet
Quoted_verified	Wenn der Wert «TRUE» ist, wird angezeigt, dass der Benutzer ein verifiziertes Konto hat.
Retweet_count	Wie oft dieser Tweet schon retweetet wurde.
Retweet_created_at	Publikationsdatum des Zitierten Beitrages mit Uhrzeit
Retweet_decription	Die benutzerdefinierte UTF-8-Zeichenkette, die das Konto beschreibt (Annullierbar) von dem der organale Tweet stammt
Retweet_favorite_count	Zeigt ungefähr an, wie oft dieser Tweet von Twitter-Nutzern gemocht wurde
Retweet_followers_count	Die Anzahl der Anhänger die das Konto, von dem der Retweet stammt, derzeit hat
Retweet_friends_count	Die Anzahl der Benutzer, denen das Konto, von dem der Retweet stammt, folgt (AKA ihre «Gefolgschaft»).

Retweet_location	Der benutzerdefinierte Standort für das Profil dieses Kontos. Dieses Feld wird vom Suchdienst gelegentlich unscharf interpretiert.
Retweet_name	Benutzername des Accounts (Statisch)
Retweet_retweet_count	Wie oft der Beitrag, von dem der Retweet stammt, schon retweetet wurde.
Retweet_screen_name	Anzeigename des Accounts (nicht Statisch)
Retweet_source	Dienstprogramm, mit dem der Tweet als HTML-formatierter String gepostet wird.
Retweet_status_id	Dieses Feld erscheint nur, wenn der Tweet ein Retweet ist. Dieses Feld enthält den ganzzahligen Wert Tweet ID des Retweets.
Retweet_statuses_count	Anzahl Beiträge die der Account, von dem das Zitat stammt, abgesetzt hat.
Retweet_text	Dieses Feld erscheint nur, wenn der Tweet ein Re-Tweet ist. Dieses Attribut enthält das Tweet-Objekt des Original-Tweets.
Retweet_user_id	Die ganzzahlige Darstellung des eindeutigen Identifikators für diesen Tweet
Retweet_verified	Wenn der Wert «TRUE» ist, wird angezeigt, dass der Benutzer ein verifiziertes Konto hat.

Tabelle 8: Variablen Beschreibung des kompletten Twitterdatensatzes

2.3.2 Twitter Hashtag der User (twitter_hashtags_MINI)

Variable Name	Beschreibung
Akteur.typ	Typ des Accounts (Person, Party, Organisation, usw.)
Akteur	Name der Organisation
Kurzel	Abkürzung des Organisationsnamens
First_Name	Vorname der Person
Last_Name	Nachname der Person
Gender	Geschlecht der Person
Year.of.Birth	Geburtsjahr der Person
Canton	Wohnkanton Kanton der Person
Zip	Postleitzahl des Wohnortes der Person
Language	Sprache
Party	Name der Mutterpartei der die Person angehört
Fraction	Fraktionszugehörigkeit im Bundeshaus, sofern es sich um eine Person handelt, die einen Sitz im Parlament innehält.
Chamber	Ratskammer in der Person Sitz inne hat
Incumbent	Bisherigen Status einer Person (Kandidaten

User_id	Die ganzzahlige Darstellung des eindeutigen Identifikators für diesen Tweet. Diese Zahl ist größer als 53 Bit und einige Programmiersprachen können Schwierigkeiten/Störungen bei der Interpretation haben. Die Verwendung einer vorzeichenbehafteten 64-Bit-Ganzzahl zur Speicherung dieses Identifiers ist sicher, wie auch die Speicherung als String.
Screen_name	Anzeigename des Accounts (nicht Statisch)
Verified	Wenn der Wert «TRUE» ist, wird angezeigt, dass der Benutzer ein verifiziertes Konto hat.
Followers_count	Die Anzahl der Anhänger, die dieses Konto derzeit hat
Friends_count	Die Anzahl der Benutzer, denen dieses Konto folgt (AKA ihre «Gefolgschaft»).
Favourites_count	Die Anzahl der Tweets, die dieser Benutzer im Laufe der Lebensdauer des Kontos gemocht hat.
Statuses_count	Anzahl Beiträge die der Account abgesetzt hat.
Party_Short	Kürzel des Parteinamens (Detailliert)
Statuses_count_period	Anzahl Nachrichten die der Account abgesetzt hat
Tot_ret_period	Anzahl Retweets aller Tweets die von dem Account stammen
Tot_fav_period	Anzahl Likes aller Tweets die von dem Account stammen
Tot_retfav_period	Anzahl aller Likes und Retweets die von dem Account stammen (Engagement Wert)
Retweet_Statuses_count_period	Anzahl Retweets die der Account gepostet hat
Quote_Statuses_count_period	Anzahl Quotes die der Account gepostet hat
Original_tweet_count_period	Anzahl selbstverfasster Tweets die der Account gepostet hat
feature	Eins der hundert häufigsten Hashtags des Accounts
frequency	Vorkommens Häufigkeit des Hashtages in absoluter Zahl des Hashtages dieses Accounts
rank	Ranghäufigkeit des Hashtags des Accounts
docfreq	Vorkommens Häufigkeit des Hashtages in den Tweets in absoluter Zahl des Hashtages dieses Accounts
group	Artefakt der DTM (Document Term Matrix)

Tabelle 9: Dieser Datensatz beinhaltet die 100 am häufigsten verwendeten Hashtag jedes Benutzers in unserm grossen Datensatz, sowie mehrere zum Teil aggregierte Kennzahlen jedes einzelnen Nutzers.

2.3.3 Twitter Statistiken (twitter_userstats_MINI)

Variable Name	Beschreibung
Akteur.typ	Typ des Accounts (Person, Party, Organisation, usw.)
Akteur	Name der Organisation
Kurzel	Abkürzung des Organisationsnamens
First_Name	Vorname der Person
Last_Name	Nachname der Person
Gender	Geschlecht der Person
Year.of.Birth	Geburtsjahr der Person
Canton	Wohnkanton Kanton der Person
Zip	Postleitzahl des Wohnortes der Person
Language	Sprache
Party	Name der Mutterpartei der die Person angehört
Fraction	Fraktionszugehörigkeit im Bundeshaus, sofern es sich um eine Person handelt, die einen Sitz im Parlament innehält.
Chamber	Ratskammer in der Person Sitz inne hat
Incumbent	Bisherigen Status einer Person (Kandidaten
User_id	Die ganzzahlige Darstellung des eindeutigen Identifikators für diesen Tweet. Diese Zahl ist größer als 53 Bit und einige Programmiersprachen können Schwierigkeiten/Störungen bei der Interpretation haben. Die Verwendung einer vorzeichenbehafteten 64-Bit-Ganzzahl zur Speicherung dieses Identifiers ist sicher, wie auch die Speicherung als String.
Screen_name	Anzeigename des Accounts (nicht Statisch)
Verified	Wenn der Wert «TRUE» ist, wird angezeigt, dass der Benutzer ein verifiziertes Konto hat.
Followers_count	Die Anzahl der Anhänger, die dieses Konto derzeit hat
Friends_count	Die Anzahl der Benutzer, denen dieses Konto folgt (AKA ihre «Gefolgschaft»).
Favourites_count	Die Anzahl der Tweets, die dieser Benutzer im Laufe der Lebensdauer des Kontos gemocht hat.
Statuses_count	Anzahl Beiträge die der Account abgesetzt hat.
Party_Short	Kürzel des Parteinamens (Detailliert)
Statuses_count_period	Anzahl Nachrichten die der Account abgesetzt hat
Tot_ret_period	Anzahl Retweets aller Tweets die von dem Account stammen
Tot_fav_period	Anzahl Likes aller Tweets die von dem Account stammen

Tot_retfav_period	Anzahl aller Likes und Retweets die von dem Account stammen (Engagement Wert)
Retweet_Statuses_count_period	Anzahl Retweets die der Account gepostet hat
Quote_Statuses_count_period	Anzahl Quotes die der Account gepostet hat
Original_tweet_count_period	Anzahl selbstverfasster Tweets die der Account gepostet hat
Sources_Vec_name	Liste der Dienstprogramme mit dem der User Beiträge gepostet hat
Sources_Vec_count	Anzahl Dienstprogramme die der User benützt hat um Beiträge zu Posten

Tabelle 10:Dieser Datensatz beinhaltet aggregierte Metadaten aller im Hauptdatensatz vorkommenden Twitter Nutzern

2.3.4 Twitter wöchentliche Statistiken (twitter_w_userstats_MINI)

Variable Name	Beschreibung
Akteur.typ	Typ des Accounts (Person, Party, Organisation, usw.)
Akteur	Name der Organisation
Kurzel	Abkürzung des Organisationsnamens
First_Name	Vorname der Person
Last_Name	Nachname der Person
Gender	Geschlecht der Person
Year.of.Birth	Geburtsjahr der Person
Canton	Wohnkanton Kanton der Person
Zip	Postleitzahl des Wohnortes der Person
Language	Sprache
Party	Name der Mutterpartei der die Person angehört
Fraction	Fraktionszugehörigkeit im Bundeshaus, sofern es sich um eine Person handelt, die einen Sitz im Parlament innehält.
Chamber	Ratskammer in der Person Sitz inne hat
Incumbent	Bisherigen Status einer Person (Kandidaten
User_id	Die ganzzahlige Darstellung des eindeutigen Identifikators für diesen Tweet. Diese Zahl ist größer als 53 Bit und einige Programmiersprachen können Schwierigkeiten/Störungen bei der Interpretation haben. Die Verwendung einer vorzeichenbehafteten 64-Bit-Ganzzahl zur Speicherung dieses Identifiers ist sicher, wie auch die Speicherung als String.
Screen_name	Anzeigename des Accounts (nicht Statisch)
Verified	Wenn der Wert «TRUE» ist, wird angezeigt, dass der Benutzer ein verifiziertes Konto hat.
Followers_count	Die Anzahl der Anhänger, die dieses Konto derzeit hat

Friends_count	Die Anzahl der Benutzer, denen dieses Konto folgt (AKA ihre «Gefolgschaft»).
Party_Short	Kürzel des Parteinamens (Detailliert)
week	Wochennummer
week_start_date	Wochenstartdatum (Montag)
week_end_date	Wochenenddatum (Sonntag)
Tot_ret_period	Anzahl Retweets aller Tweets die von dem Account stammen
Tot_fav_period	Anzahl Likes aller Tweets die von dem Account stammen
Tot_retfav_period	Anzahl aller Likes und Retweets die von dem Account stammen (Engagement Wert)
Retweet_Statuses_count_period	Anzahl Retweets die der Account gepostet hat
Quote_Statuses_count_period	Anzahl Quotes die der Account gepostet hat
Original_tweet_count_period	Anzahl selbstverfasster Tweets die der Account gepostet hat

Tabelle 11: Dieser Datensatz beinhaltet wöchentlich aggregierte Metadaten aller im Hauptdatensatz vorkommenden Twitter Nutzern

2.3.5 Twitter Friendslist (twitter_friendslist_MINI)

Diese Datei enthält eine verschachtelte Liste von den Freundeslisten aller 1284 für diese Analyse beobachteten Accounts, welche für die Netzwerkanalyse verwendet wird. Diese Datei ist nur als RDS-Datei erhältlich, da es sich um eine Netzwerkliste handelt mit Objekten und nicht um einen einfachen Datensatz mit Zeilen und Spalten.

2.4 Überblick Facebook Daten

2.4.1 Facebook Pages (facebook_userstats_MINI)

Variable Name	Beschreibung
Profil	Profilname
wachstum.absolut	Absolutes Wachstum der Facebook Page
posts.pro.tag	Anzahl durchschnittliche Posts pro Tag
year_of_birth	Geburtsjahr des Kandidaten/ der Kandidatin
fans	Anzahl Personen, die der Page folgen
anzahl.likes	Anzahl Likes, die die Page hat
anzahl.posts	Anzahl Posts der Page
engagement	Engagement Wert der Seite (Erfolgswert / Reichweite))
woechentliches.wachstum	Wöchentliches Wachstum dieser Page
page.performace.index	Page Performance Index von dieser Woche
post.interaktion	Anzahl Interaktionen
gesamtzahl.reaktionen.kommentare.shares	Summe aller Reaktionen, Kommentare und Shares zusammen
anzahl.kommentare	Anzahl Kommentare
gewichtetes.engagement	Gewichtetes Engagement dieser Page
taegliches.wachstum.in.	Durchschnittliches Wachstum der Page in Prozent pro Tag in dieser Woche
wachstum.absolut.pro.tag	Durchschnittliches Wachstum der Page pro Tag in dieser Woche
reaktionen.pro.post	Summer der durchschnittlichen Reaktionen pro Post in dieser Woche
shares.pro.post	Summer der Reaktionen pro Post in dieser Woche
summe.der.reichweite.einzelner.posts	Summer der Reichweite aller Posts dieser Woche
summe.der.impressionen.einzelener.posts	Summe der Impressionen aller Posts dieser Woche
gesamtreichweite.organisch	Organische Reichweite der einzelnen Posts in dieser Wiche
gesamtreichweite.pro.tag	Durchschnittliche Reichweite pro Tag in dieser Woche
seitenaufrufe.pro.tag	Durchschnittliche Anzahl Seitenaufrufe pro Tag in diese Woche
gesamtreichweite	Totale Reichweite der Page in dieser Woche
durchschnitlliche.reichweite.einzeler.posts	Durchschnittliche reichweiter der einzelnen Posts dieser Woche
seitenaufrufe	Anzahl Aufrufe der Seite in dieser Woche
startDate	Wochenstartdatum (Aggregationslevel Woche)
endDate	Wochenenddatum (Aggregationslevel Woche)
party	Parteiabkürzung bei der die Person Mitglied ist

Tabelle 12: Überblick über den Datensatz mit allen Informationen zu den Facebook Pages der Kandidatinnen und Kandidaten

2.4.2 Facebook Beiträge (facebook_class_sent_MINI)

Variable Name	Beschreibung
Profil	Profilname
date	Publikationsdatum des Beitrages auf der Page
time	Uhrzeit der Publikation des Beitrages auf der Page
txt	Text des Beitrages
nLikes	Anzahl Likes, die der Beitrag erhalten hat
nComments	Anzahl Kommentare, die der Beitrag erhalten hat
interAct	Anzahl Interaktionen, die der Beitrag generiert hat
party	Parteiabkürzung bei der die Person Mitglied ist
la	Sprache des Textes
sentiment_value	Tonalitätswert des Textes
positive_words	Alle Positiven Schlüsselwörter im Text
negative_words	Alle negativen Schlüsselwörter in Text
selectsclass	Thema des Textes klassifiziert mit dem Twitter Algorithmus

Tabelle 13: Überblick über den Datensatz mit allen Beiträgen der Pages von den Kandidatinnen und Kandidaten auf Facebook

3 Technischer Bericht

3.1 Kurzanleitung der automatisierten Medienanalyse

Die Hauptaufgabe der Selects Medienanalyse von 2019 besteht aus dem zusammentragen der Daten von den verschiedenen Quellen, dem Vorverarbeiten der einzelnen Artikel und Beiträge, dem Abspeichern der Daten in Echtzeit, sowie der automatisierten Identifikation der relevanten Dokumente und Nachrichten (Klassifikation), der Extraktion oder Annotation des Themas (Klassifikation), Partei (Annotation), Person (Named Entity Recognition) und das berechnen der Tonalität jedes einzelnen Dokumentes aller beobachten Quellen. Das verwendete Framework bedient sich an unserer Datenverarbeitungsinfrastruktur, welche ein Daten Analyse Framework beinhaltet.

Dieses umfasst ein vollautomatisiertes Ingestion System über mehrere Nodes, die von verschiedenen Quellen Daten herunterladen und in Echtzeit verarbeiten und auf eine Datenbank schreiben und sichern. Darüber hinaus beinhaltet unser Framework auch die Möglichkeit Daten automatisiert in der Datenbank zu verarbeiten. Dies gestattete es uns

die Daten vollautomatisiert mit trainierten Algorithmen und Verfahren täglich zu klassifizieren, was es uns erlaubt immer mit aktuellen Daten zu Arbeiten ohne ständiges manuelles Nachcodieren neuer Artikel von bestehenden und neuen Quellen.

3.2 Framework

3.2.1 Datenbeschaffung

Suchen, Laden und vorbereiten der Daten von allen Quellen für das Hochladen in die Datenbank. Hierbei wurden verschiedene manuell erstellte Listen von den Kandidierenden und anderen AkteurInnen benutzt, um die Daten mit Sinnvollen Metadaten anzureichern.

3.2.2 Tonalität

Berechnung der Tonalität (Sentiment) aller gesammelten Texte basierend auf einem mehrsprachigen Wörterbuch verfahren, welches von Proksch et al. (2019) für politische Texte entwickelt wurde und an Reden des Europäischen Parlamentes getestet wurde.

3.2.3 Klassifikation

Verwendung von mit H2O.ai vortrainierten Ensemble Modellen zur Klassifikation von Zeitungsartikeln und Beiträgen aus den sozialen Medien in verschiedene politische Themengebiete.

3.2.4 Named Entity Recognition

Wörterbuch basierende Named Entity Recognition die Zeitungsartikel nach Politikern und Vorwahlbefragungen mit einem Bash Regex Skript durchsucht und jegliche Treffer abspeichert.

3.2.5 Zusätzliche Schritte

- Trainings Daten: Ein separater Workflow wurde verwendet, um Trainingsdaten für die Textklassifikation zu erstellen. Dies Umfasst ein Matching Prozess der Klassifizierte APS Zeitungsartikel (Année politique Suisse) mit den SMD Texten verbindet und unser Klassifikation Schema aus den APS Themen herausbildet. Dies ermöglichte es uns Texte nicht nur nach Relevanz zu klassifizieren, sondern auch nach Themengebieten akkurat zu klassifizieren. Darüber hinaus umfasst dieser Teil ebenfalls das Erstellen eines Trainingsdatensatzes aus Twitter

Textnachrichten mit Themen orientierten Hashtags zum Trainieren eines Algorithmus der sowohl Twitter Nachrichten als auch Beiträge, die auf Facebook Pages abgesetzt wurden, klassifizieren kann.

- SMD Ensemble Training: Das Testen und Trainieren eines Ensemble Klassifikationsalgorithmus für die Klassifikation von Zeitungsartikel bestehend aus mehreren Skripten die verschiedenen Algorithmen mit Random Grid-Search trainieren, die später im Ensemble Algorithmus verwendet werden. Sowie dem konstruieren eines binären Wörterbuch basierten Klassifikationsalgorithmus, der die Texte in einem ersten Schritt als politisch oder nicht politisch einstuft.
- Twitter / Facebook Ensemble Training: Das Testen und Trainieren eines Ensemble Klassifikationsalgorithmus für die Klassifikation von Beiträgen in den sozialen Medien bestehend aus mehreren Skripten die verschiedenen Algorithmen mit Random Grid-Search trainieren, die später im Ensemble Algorithmus verwendet werden.

3.3 Ingestion System

3.3.1 SMD (Schweizerische Mediendatenbank)

Der Download der Daten erfolgt mithilfe eines Python Skriptes, welches die Elasticsearch Datenbank der SMD nach Zeitungsartikel von einer definierten Liste von Zeitungen abfragt. Die Abfrage wird täglich ausgeführt und fragt die Datenbank jeweils nach allen Zeitungsartikeln vom vorgestrigen Tag ab aus der Liste ab.

Die einzelnen Artikel werden als json-Datei von der Datenbank zur Verfügung gestellt und von uns auf unserer Infrastruktur zwischengespeichert und in ein angemessenes Format transformiert. Hierbei haben wir sichergestellt, dass alle Artikel alle Metadaten aufweisen, die wir für unsere Analysen benötigen und diese in einer einheitlichen Form vorliegen. Anknüpfend lädt ein Skript alle neuen Artikel auf unsere eigene Elasticsearch Datenbank. Die Datenbank wird täglich abgefragt, um das Downloadlimit von 10’000 Artikeln pro Abfrage nicht zu überschreiten.

Die Abfrage erfolgt jeweils für den vorgestrigen Tag, da die Datenbank der SMD oft erst nach zwei Tagen vollständige Daten enthält. Hiermit stellen wir sicher, dass die Daten vollständig sind und es zu keinem Download Fehler kommt.

3.3.2 Twitter

Für Twitter benutzen wir die offizielle API (Application Programming Interface) zum Herunterladen der Nachrichten. Dies gestattet es uns auch sehr viele Nachrichten in kurzer Zeit zu erhalten mit allen Metadaten, die Twitter zur Verfügung stellt.

Damit wir möglichst alle Twitter Nachrichten zu den Schweizer Wahlen erhielten verwendeten wir zwei verschiedene Skripts, die beide die REST API (Representational State Transfer API) verwenden, da wir nur bei dieser Schnittstelle sicher gehen können, dass wir jeweils alle Nachrichten erhalten und nicht nur eine unvollständige Stichprobe wie die bei der Stream API der Fall sein kann.

Für das eine Skript laden wir jeweils alle neuen Nachrichten von einer Liste von definierten Nutzern herunter, die von uns aus der offiziellen Kandidierendenenliste von Smartvote erstellt wurde, sowie zusätzlichen wichtigen institutionellen NutzerInnen auf Twitter, wie Zeitungen, Parteien, Verbänden und die offiziellen Accounts des Bundes selbst. Dieses Skript wird täglich automatisch ausgeführt und lädt jeweils die letzten 2000 Nachrichten eines jeden Nutzers herunter und aktualisiert schon vorhandene Nachrichten im Archiv. Hiermit ist es auch ausgeschlossen, dass wir eine Nachricht nicht herunterladen. Jeder einzelne Tweet wird hierbei als json Datei im Archiv abgespeichert. Anschliessend werden neue Nachrichten mit einem Skript in die Datenbank übertragen und alte Nachrichten werden jeweils einmal pro Woche von einem anderen Skript mit den aktualisierten Metadaten in der Datenbank versehen. Damit wird sichergestellt, dass alle Nachrichten auch wirklich die plausiblen Zahlen bezüglich Likes und dergleichen enthalten.

Ein zweites Skript bezieht täglich von der REST API alle Nachrichten die gewisse Hashtags (#) enthalten, die von uns ausgewählt werden, da diese sehr häufig in Nachrichten vorkommen, die für die Wahl und den Wahlkampf als auch für die SRF Arena relevant sind. Diese werden dann jeweils auch im Archiv abgespeichert.

Da diese Nachrichten zum Teil auch Nachrichten enthalten, die von den separat mitgehörten Accounts stammen, werden die Daten jeweils mit zwei Tagen Verzögerung in die Datenbank geschrieben. Dies erlaubt es uns auf einfache Weise Nachrichten, die schon in der Datenbank sind vor dem Hochladen herauszufiltern. Damit wird sichergestellt, dass keine Daten doppelt in die Datenbank geschrieben werden können

und dies führt auch dazu, dass die Nachrichten jeweils von dem Account basierten Skript stammen, da dieses durch die Smartvote Liste zusätzliche Informationen über die Verfasser in die Datenbank schreiben kann, wie zum Beispiel die Parteizugehörigkeit eines Users. Die Daten liegen also immer mit den umfassendsten Metadaten in der Datenbank vor.

3.3.3 Facebook

Wie schon eingangs bei den Datenquellen erwähnt verwenden wir eine externe Quelle, um die Beiträge von Facebook Pages herunterzuladen. Die Plattform von FanpageKarma ermöglicht es uns zumindest alle Pages von Kandidaten zu überwachen.
Dabei kann selbst FanpageKarma nur Daten von öffentlichen Pages zu sammeln, weshalb wir nur von den Kandidierenden Daten sammeln konnten, die eine Page besitzen. Die Daten, die von FanpageKarma gesammelt werden, haben wir am Ende mit einem teilautomatisierten Webscraper heruntergeladen und für unsere Verwendungszwecke vorbereitet. Wir mussten auf einen simplen Webscraper zurückgreifen, der für jeden einzelnen Account separat gestartet werden muss, da es nicht möglich ist mehrere Pages zusammen zu scrapen, da die Webseite von FanpageKarma dies nicht hinbekommt ohne, abzustürzen. Zudem erschwert die Zweifaktorauthentifizierung von Facebook das ganze Prozedere noch einmal ein wenig, da beim starten des Scrapers jedes Mal das Login betätigt werden muss. Trotz dieser Widrigkeiten funktioniert es die Daten schlussendlich Zeitschonend herunterzuladen.

3.4 Tonalität

Bei der Berechnung der Tonalität verwenden wir eine wörterbuchbasierte Methode, die von Proksch et al. (2019) vorgestellt wurde. Es handelt sich um eine der neuesten Methoden dieser Art zur Bestimmung der Tonalität.
Das Hauptargument, weshalb wir dieses Verfahren nutzen, ist dessen Fähigkeit die Tonalität nicht nur in einer Sprache zu berechnen, sondern dies in verschiedenen Sprachen zu bewältigen mit annähernd denselben Resultaten für gleiche Texte in unterschiedlicher Sprache. Diese Eigenschaft ist eine Voraussetzung für das Berechnen der Tonalität von Schweizer Zeitungsartikel und Beiträgen in den sozialen Medien da wir

drei grosse Landesprachen (Deutsch, Französisch, Italienisch) haben und noch eine weitere kleine (Rätoromanisch). Letztere können wir aber nicht berücksichtigen, denn für diese gibt es kein Wörterbuch und wir sammeln auch keine Daten in dieser Sprache.
Darüber hinaus eignet sich das Verfahren von Proksch et al. (2019) besonders für politische Texte, da sie ihre Methode für diese Textart optimiert und ausgiebig an Texten des europäischen Parlamentes getestet haben. Die von Proksch et al. (2019) gezeigte Übereinstimmung des Sentiments ihrer Methode und Handkodierter Texte ist mit mehr als 80 % in allen von uns benutzten Sprachen (Deutsch, Französisch und Italienisch) mehr als Ausreichend um die Tonalität wiedergeben zu können. Daher können wir mit grosser Sicherheit davon ausgehen, dass die politischen Texte für die Selects Medienanalyse eine Tonalität erhalten, die sicherlich gehaltvoll für politische Zeitungsartikel ist.
Der Ansatz von Proksch et al. (2019) basiert auf dem bekannten und verbreiteten Wörterbuch Lexicoder von Young und Soroka (2012). Dieses wurde von Proksch et al. (2019) mit einer optimierten automatisierten Übersetzungsmethode in verschiedene Sprachen übersetzt, um die Tonalität auch in anderen Sprachen erfassen zu können. Proksch et al. (2019) argumentieren für das übersetzen des Wörterbuches statt der Texte aus zwei Gründen.
Erstens lässt sich ein validiertes Wörterbuch zur Messung der Tonalität heute mit den verfügbaren Übersetzungsservices sehr leicht und mit hoher Qualität übersetzen. Darüber hinaus ist es mit weiterem Aufwand auch noch möglich die Validität der Übersetzung zu prüfen und es kann nachträglich noch manuell annotiert und verbessert werden. Im Gegensatz dazu ist es viel aufwändiger viele Texte mit diesen Tools zu übersetzen. Es benötigt einfach mehr Zeit und die Übersetzungsqualität ist dann auch nicht mehr ganz so gut, da diese Tools immer noch Mühe haben den Kontext zu übersetzen, was es dann viel schwerer macht, die Tonalität zu messen. Zudem lässt sich die Validität der Übersetzungen ganzer Texte nicht mehr so einfach prüfen. Die Forscherin verlässt sich hierbei stark auf die angegebene Validität und Reliabilität der verwendeten Services.
Das zweite Argument für die automatische Übersetzung eines getesteten Wörterbuchs ist der enorme Zeitgewinn. Trotz der heutigen Fähigkeiten von Computern grosse Mengen an Text schnell zu übersetzen und dann die Tonalität zu berechnen dauert es immer noch

viel länger als das Wörterbuch zu übersetzen und dann jeweils alle Texte in ihrer Originalsprache nach ihrer Tonalität zu untersuchen. Zudem ermöglicht diese Methode sprachspezifisch manuell angepasste und getestete Wörterbücher für die Kalkulation der Tonalität.

Unsere Methode zur Berechnung der Tonalität orientiert sich sehr stark an der von Proksch et al. (2019). Bei der Kalkulation wird zuerst der Text maschinenlesbar gemacht. Hierfür entfernen wir zuerst jegliche Punktationen, Zahlen und URLs in den Texten, sowie alle Zeilenumbrüche und andere Artefakte. Danach wird die Sprache festgestellt. Dies geschieht entweder über die Sprachvariable in den Metadaten oder durch Googles Compact Language Detector (Version 2) der die Sprache des Textes erkennen kann. Anschliessend wird die Tonalität des Textes berechnet. Die Berechnung der Formel von Proksch et al. (2019) sieht wie folgt aus:

$$Tonalität = \log\left(\frac{pos + 0.5}{neg + 0.5}\right)$$

Das Wörterbuch dient also dazu alle negativen und alle positiven Begriffe in jedem Text zu finden und sie zur Berechnung der Tonalität zu zählen.

Bei dieser Formal muss vor allem bei kurzen Texten mit einer neutralen Tonalität mit dem Wert 0 Acht gegeben werden. Dies entsprich dann oftmals einfach keinem positiven oder negativen Begriff im Text. Bei langen Texten bedeutet eine neutrale Null hingegen, dass es genau gleich viele negative, wie positive Begriffe im Text hat und er hiermit ausgewogen hinsichtlich seiner Tonalität ist. Anzumerken ist hier, dass Young und Soroka (2012) feststellten, dass Texte öfters positiver als negativ sind.

Aus diesen Resultaten geht hervor, dass Zeitungsartikel mit dieser Methode sicherlich gut analysiert werden können. Denn deren Texte sind lange genug womit es sehr selten keinen positiven oder negativen Begriff im Text haben wird. Bei den Beiträgen aus den sozialen Medien sieht es etwas anders aus. Hier gibt es wegen der kurzen Texte häufig Resultate mit einer neutralen Tonalität, bei der kein Begriff vorkommt, der eine Wertung der Tonalität zulässt. Nichtsdestotrotz funktioniert das berechnen der Tonalität gut, sobald ein Begriff gefunden wird, da diese kurzen Beiträge klar formuliert werden müssen. Wir haben dies anhand einer kleinen Stichprobe manuell überprüft und ausgewertet. In der Tabelle 14ist zu sehen, dass die Tonalität auch relativ gut berechnet werden konnte für kurze Beiträge aus den sozialen Medien:

Sentiment	Präzision	Sensitivität	F1-Mass
Positiv	0.920	0.643	0.757
Negativ	0.667	0.608	0.636
Neutral	0.421	0.818	0.556

Tabelle 14: Evaluationstabelle für die Tonalität (Sentiment) der Beiträge aus den sozialen Medien

Die Klassifikation funktioniert im Allgemeinen recht gut, die tiefe Präzision bei den neutralen Textnachrichten liegt alleine daran, dass nur eine Null als Neutral gilt, es aber recht häufig bei der Berechnung vorkommt, dass die Werte auch mal leicht positiv oder negativ sein können. Wenn man die Grenze nicht so genau zieht sieht es viel besser aus. Das einzige wirkliche Problem bei der Berechnung der Tonalität tritt bei Texten auf, die Sarkasmus beinhalten. Diese Texte weisen in den meisten Fällen eine falsche Tonalität auf.

3.5 Klassifikation

Die Klassifikation der Texte basiert auf sehr ähnlichen Verfahren sowohl für die Zeitungsartikel wie auch für Beiträge aus den sozialen Medien. Der Hauptunterschied ist die zweistufige Klassifikation der Zeitungsartikel gegenüber den Texten von sozialen Medien, da wir nicht nur politisch relevante Artikel der Schweiz bekommen, sondern auch noch alle anderen Artikel der jeweiligen Zeitungen.

Deshalb entscheidet ein Entscheidungsbaum zuerst, ob es sich um einen politisch relevanten Text in der Schweiz handelt oder nicht. Erst anschliessend wird der Ensemble Algorithmus verwendet, um einem Text ein Thema zuzuordnen.

Bei den Beiträgen aus den sozialen Medien verwenden wir direkt einen Ensemble Klassifikationsalgorithmus, der jeden Beitrag klassifiziert, wobei die nicht politischen Artikel und nicht klassifizierbaren anderen Beiträge in einer Klasse landen, da wir zu diesem Zweck Trainingsdaten generieren konnten im Gegensatz zu den Zeitungsartikeln bei denen dies nicht gemacht wurde.

Die verwendeten Ensemble Methoden bestehen für die Zeitungsartikel in Deutsch und Französisch aus jeweils zwei *Deep-Learning Modellen* und einem *Gradient-Boosting Modell.* Das *Random-Forest Modell* wurde nicht verwendet, da es im Ensembleverfahren keine Verbesserung bei der Klassifikation mit sich bringt. Im Ensemble Modell werden

die Wahrscheinlichkeiten für jedes Thema ausgeben und dann die Wahrscheinlichkeiten aufsummiert. Der Text wird demjenigen Thema zugeordnet, das die höchste Wahrscheinlichkeit über alle Algorithmen aufweist. Für das Ensemble Modell der sozialen Medien verwenden wir dasselbe Schema ebenso mit zwei *Deep-Learning Modellen* und einem *Gradient-Boosting Modell*. Wir verzichten auch hier auf das *Random-Forest Modell*.

3.5.1 Binärklassifikation (nur für SMD)

Die binäre Klassifikation ist ein einfacher Entscheidungsbaum, der für jeden Text zuerst berechnet, ob mehr Schlüsselwörter aus der Wortliste für Schweizer Politik gefunden wurden oder aber für die Liste mit nicht-politischen Begriffen. Falls mehr Wörter aus der Wortliste für Schweizer Politik gefunden werden, wird dem Text die Klasse Schweizer-Politik zugewiesen.

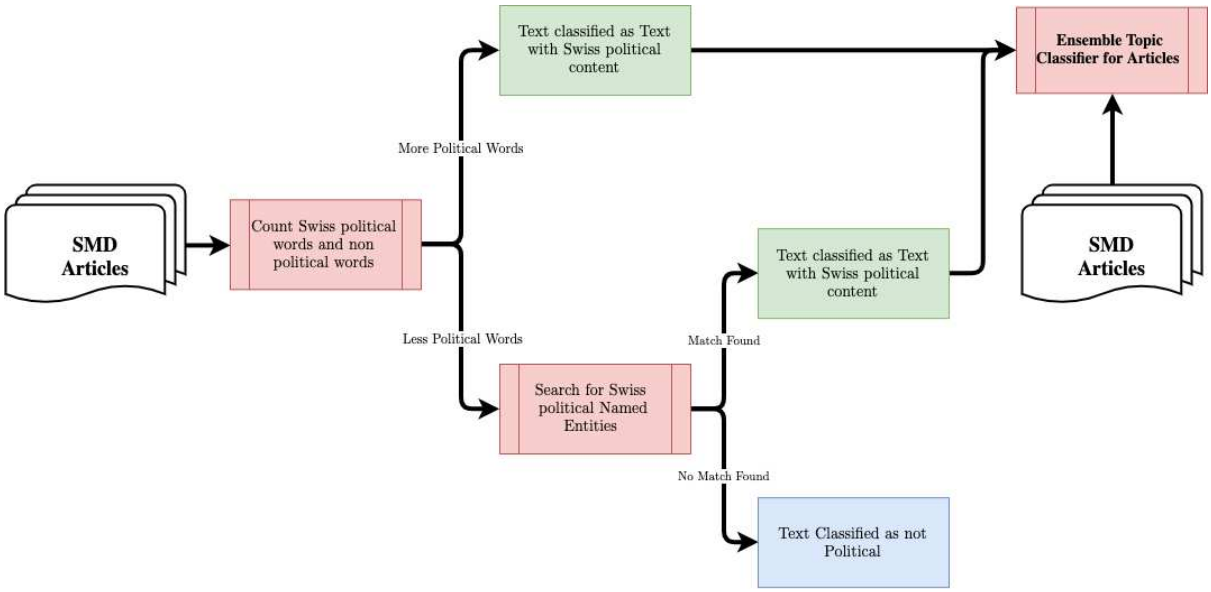

Abbildung 1: Binäres Klassifikationsschema für die Zeitungsartikel aus der SMD Datenbank

Trifft der andere Fall ein, werden vom Text die Eigennamen durchsucht. Finden sich hierbei Namen aus unserer Eigennamenliste (Named Entities), die aus Kandidierendennamen, Parteinamen, Organisationsnamen und anderen Namen besteht, wird der Text schlussendlich auch als Schweizer Politik klassifiziert. Nur wenn auch in diesem Schritt nichts Gefunden wird, wird der Text als nicht-politisch klassifiziert.

Für die manuelle binäre Klassifikation benötigen wir generell eine Liste von Wörtern die häufig und nur in Artikeln vorkommen, die mit Schweizer Politik zu tun haben, sowie eine weitere Liste mit Wörtern die sehr häufig in Texten vorkommen die nichts mit Schweizer Politik zu tun haben. Diese zwei Listen mussten von Hand erstellt werden. Für die erste Liste verwenden wir Wörter wie Referendum, Nationalrat, Bundesrat und Abstimmung, da diese häufig mit Schweizer Politik in Verbindung stehen.
Für die andere Liste verwenden wir vor allem Wörter, die mit Sport, Kunst und Unterhaltung zusammenhängen. Darüber hinaus haben wir eine Eigennamenliste erstellt, die alle Namen der Kandidierenden der Wahlen von 2019 umfasst, sowie der Ehemaligen und bisherigen Ratsmitglieder des Ständerates und des Nationalrates, hohe Beamte, Parteinamen, Organisationsnamen, Departemente, Behördennamen und Abstimmungstitel umfasst.

3.5.2 Trainingsdaten

Bevor ein Klassifikationsalgorithmus trainiert werden kann, bedarf es geeigneter Trainingsdaten. Hierfür erstellten wir mit Hilfe von APS (Anée Politique Suisse) jeweils einen Trainingsdatensatz für Zeitungsartikel auf Deutsch und Französisch. Für das Erstellen der Trainingsdaten war das Matching der APS Daten mit den SMD Daten über den Zeitungsnamen, das Publikationsdatum und ein Keyword Matching aus dem Titel der SMD Artikel nötig. Von den APS Daten wurden die Dateinamen genutzt, welche das Publikationsdatum, Zeitungsname, ein Keyword aus dem Titel und das Thema nach APS enthalten, anstatt auf die PDFs der Artikelseite zurückzugreifen. Ohne die Daten von der APS wäre es nicht möglich gewesen einen Datensatz zu erstellen, der schon klassifizierte Texte enthält.
Für Italienisch konnten wir keinen Trainingsdatensatz erstellen, da hier keine klassifizierten Textdateien erstellt werden konnten aus dem Matching der APS Daten und den Zeitungsartikel der SMD, da die einzige italienische Zeitung in der APS Datensammlung der Corriere del Ticino ist. Dieser ist aber leider nicht Teil der SMD Daten.
Die Trainingsdaten für Deutsch und Französisch enthalten alle Artikel die eindeutig zwischen SMD und APS Daten identifiziert werden konnten von Januar 2012 bis

September 2019. In der Tabelle 15 sind die wichtigsten Zahlen zu diesen Trainingsdaten ersichtlich.

Themen	Anzahl Artikel (De)	Anteil der Artikel (De)	Anzahl Artikel (Fr)	Anteil der Artikel (Fr)
Total Anzahl Artikel	137852	100%	20855	100%
Politisches System	27096	19.66%	4538	21.76%
Wirtschaft	16582	12.03%	1732	8.30%
Öffentliche Dienste / Infrastruktur	14637	10.62%	1737	8.33%
Erziehung & Kultur	12096	8.77%	1272	6.10%
Umwelt & Energie	9889	7.17%	1157	5.55%
Gesundheitswesen	7244	5.25%	1388	6.66%
Sozialversicherung & Sozialstaat	7072	5.13%	1367	6.55%
Recht & Ordnung	6886	5%	823	3.95%
Immigration & Asyl	6707	4.87%	1460	7%
Finanzen & Steuern	5648	4.10%	1151	5.52%
Internationale Beziehungen & Armee	5589	4.05%	1179	5.65%
Landwirtschaft	3997	2.90%	466	2.23%
Regionen & Nationaler Zusammenhalt	3119	2.26%	457	2.19%
EU / Europa	3072	2.23%	878	4.21%
Arbeitsmarkt	2928	2.12%	606	2.91%
Nicht klassifizierbar	2359	1.71%	164	0.79%
Andere Probleme	1721	1.25%	162	0.78%
Geschlechterfragen & Diskriminierung	1210	0.88%	318	1.52%

Tabelle 15: Zusammensetzung der Trainingsdaten für die Zeitungsartikel

Bei den Trainingsdaten zur Klassifikation der Beiträge aus den sozialen Medien haben wir die Trainingsdaten selbst konstruiert. Dafür haben wir die Nachrichten auf Twitter mit Hilfe der 2'000 häufigsten Hashtags im Korpus, jeweils auf Deutsch und Französisch, zur Klassifikation von acht verschiedenen politischen Themen benutzt und eine neunte Klasse als Nicht klassifizierbar erstellt.

Von den 2000 verschiedenen Hashtags auf Deutsch und Französisch konnten wir nur acht Themen herausfiltern mit ca. 300 Hashtags, die einer der acht Klassen zugewiesen werden konnten. Die anderen ca. 1700 Hashtags wurden nicht verwendet für den Trainingsdatensatz, da sie zu weitläufig und zu diffus sind, um Themen aus ihnen zu bilden. Die nicht klassifizierbaren Tweets im Trainingsdatensatz wurden in beiden Sprachen mit Tweets trainiert, die von Accounts stammen, die Sportklubs gehören und

ein paar InfluencerInnen und SportlerInnen. Dies verbessert die Falschpositive Klassifikation von den neun Themen erheblich. In der Tabelle 16 sind die wichtigsten Zahlen zu den Trainingsdaten für die sozialen Medien ersichtlich.

Themen	Anzahl Artikel (De)	Anteil der Artikel (De)	Anzahl Artikel (Fr)	Anteil der Artikel (Fr)
Total Anzahl Tweets	83146	100%	11277	100%
Wahlen	30752	36.99%	5393	47.82%
Umwelt & Energie	13441	16.17%	1957	17.35%
EU / Europa	9509	11.44%	663	5.88%
Nicht klassifiziert	6580	7.91%	908	8.05%
Immigration & Asyl	6398	7.69%	345	3.06%
Geschlechterfragen & Diskriminierung	5335	6.42%	684	6.07%
Sozialversicherung & Sozialstaat	4557	5.48%	682	6.05%
Abstimmung	2812	3%	47	0.42%
Finanzen & Steuern	2686	3.23%	298	3%
Gesundheitswesen	1076	1.29%	300	2.66%

Tabelle 16: Zusammensetzung der Trainingsdaten für die sozialen Medien

3.5.3 Feature Engineering

Damit ein Text überhaupt von einem Algorithmus klassifiziert werden kann, muss dieser nummeriert werden. Hierfür verwenden wir statt einer einfachen Wortmatrize (document term matrix) ein Word2Vector Modell, um aus vielen Texten ein Wortmodell zu bauen, dass für die Klassifikation verwendet werden kann. Der Vorteil eines Word2Vector Modelles über eine DTF oder TF-IDF (term frequency inverse document frequency) ist, dass statt der reinen Gewichtung der verschiedenen Wörter in den Texten auch der Kontext zu einem bestimmten Grad rekonstruiert werden kann. Das Word2Vector Modell besteht aus einem zweischichtigen neuronalen Netzwerk, dass aus einem grossen Textkorpus einen Vektorraum von mehreren hundert Dimensionen erstellt, wobei jedem Wort im Korpus eine bestimmte Stelle in diesem Vektorraum zugewiesen wird. Dabei ist die Positionierung davon abhängig, dass Wörter, die in einem ähnlichen Kontext vorkommen, nahe beieinander im Vektorraum stehen. Da wir keine Wörter benutzen können, um das Neuronale Netzwerk zu trainieren, wird ein Wort durch einen Vektor des gesamten Vokabulars repräsentiert, das nur eine Eins an der Stelle aufweist, die dieses Wort repräsentieren soll (one-hot-vector). Unser Word2Vector

Modell verwendet die Skip-Gram Methode, um Wörter vorherzusagen, die in einem ähnlichen Kontext stehen, wie das Zielwort. Die Skip-Gram Methode behandelt jedes einzelne Wortpaar als eine neue Observation von *n* Wörtern um das Ziel Wort herum, was bei grossen Datensätzen sehr effizient ist, da es bei einem sehr grossen Korpus nicht mehr möglich ist oder sehr aufwendig wird alle möglichen Wortkombinationen zu modellieren (Guthrie et al. 2006, Mikolov et al. 2013). Der Trick des Word2Vector Modelles ist es wie bei jedem anderen neuronalen Netzwerk die Gewichte der einzelnen Neurone so lange anzupassen bis die Gewichte die kleinste mögliche Verlustfunktion erreichen. Nun wird dieses Netzwerk aber nicht verwendet, um eben diese Funktion für Vorhersagen zu benutzen, sondern es werden die einzelnen Gewichte der Neuronen verwendet, um den Vektorraum der Worteinbettungen zu repräsentieren. Das heisst wir berechnen die Wahrscheinlichkeit eines Wortes, dass es neben einem bestimmten Wort vorkommt. Dies macht es dem Word2Vector Modell möglich verschiedene Ähnlichkeiten zwischen Wörtern wiederzugeben, wie die Semantik und die Syntax. Der Nutzen hiervon ist, dass für eine Klassifikation von Textdaten davon ausgegangen werden kann, dass ähnliche Wörter im Word2Vector Modell öfters auch auf ein ähnliches Thema hindeuten werden, was den Klassifikationsalgorithmus verbessern sollte. Zur Verbesserung des Word2Vector Modelles haben wir die Texte standardisiert. Das heisst wir haben Stopp-Wörter (der, die, das, und, usw.), Zahlen und Satzzeichen entfernt, da sie nur wenig zum Informationsgehalt der Texte beitragen. Damit wir ein möglichst effizientes schnelles Word2Vector Modell bekommen, haben wir verschiedene Modell mit unterschiedlicher Dimensionalität getestet, um zu sehen ab welcher Dimensionalität sich das Modell nicht mehr positiv auf den Klassifikationsalgorithmus auswirkt. Dabei haben wir festgestellt, dass ab einer Dimensionalität von mehr als 600 die Qualität nicht mehr nennenswert ansteigt, weshalb wir uns auf Modelle von 600 Neuronen festgelegt haben. Diese Observation wurde vor allem von Pennington et al. (2014) in ihrer Arbeit zur Erstellung eines umfassenden Modelles zur Erfassung von feinen semantischen und syntaktischen Gesetzmässigkeiten in Wortrepräsentationen durch verschiedene Word2Vector Modelle untermauert.

3.5.4 Ensemble Training

Damit wir mit den Trainingsdaten aus Zeitungsartikel und Beiträge aus den sozialen Medien zuverlässig klassifizieren konnten, entschieden wir uns für vier Ensemble Algorithmen. Zwei für die Daten der SMD und zwei für die Beträge aus den sozialen Medien, jeweils einmal für Deutsch und Französisch. Die Verwendung einer Ensemble-Methode, die aus mehreren verschiedenen Klassifikationsmodellen besteht, beruht auf der Erkenntnis, dass es mit Ensemble Methoden möglich ist, die Stärken mehrere Algorithmen miteinander zu verbinden und so deren Schwächen zu mindern.

Ein Ensemble Algorithmus nutzt eine bestimmte Anzahl von verschiedenen oder ähnlichen Algorithmen, um bessere Ergebnisse zu erhalten als mit einem einzelnen Algorithmus. Die Berechnung der Ergebnisse dieser Menge von Algorithmen dauert zwar länger als die Auswertung eines einzelnen Algorithmus, allerdings kann mit einer viel geringeren Rechentiefe ein in etwa gleich gutes Ergebnis erreicht werden, wie mit einem sehr komplexen rechenintensiven einzelnen Algorithmus. Unser Ensemble Modell verwendet für die schlussendliche Klassifikation die Summe aller Wahrscheinlichkeiten der einzelnen Klassifkationsthemen und wählt dann die Klasse mit der grössten Wahrscheinlichkeit.

Die Klassifikationsalgorithmen wurden mit H2O trainiert und in R implementiert. Zum Training der Modelle wurden alle Klassifikationsalgorithmen mit einer Random-Grid Suche mit einem umfassenden Hyperparameterraum optimiert, dies geschah über jeweils mindestens 24 Stunden (Bergstra and Bengio 2012). Diese Suchmethode ermöglicht es relativ viele verschiedene Modellkonfigurationen in kurzer Zeit zu trainieren. Dies führt erheblich schneller zu guten Modellen, als mit einer klassischen Grid Suche, die den ganzen Hyperparameterraum absucht, um eine genügend grosse Anzahl an diversen Modellkonfiguration auszuprobieren.

Für die Klassifikation der Zeitungsartikel der SMD entschieden wir uns für eine Ensemble-Methode, die jeweils die zwei besten Deep-Learning Modelle und das beste Gradient-Boosting Modell verwendet. Der Grund für diese Modellauswahl in den Ensemble Modellen liegt an der allgemeinen Modellgüte der verschiedenen Algorithmen. Bei den Zeitungsartikeln weisen die besten Deep-Learning Modelle jeweils eine Genauigkeit von mehr als 75 % auf, gefolgt von den besten Gradient-Boosting Modellen, die eine Genauigkeit von knapp 70 % erreichen. Es stellt sich hier natürlich die Frage,

warum überhaupt ein Ensemble Modell verwendet wird, wenn die Deep-Learning Modelle über 75 % korrekt klassifizieren. Der Grund ist, dass die Fehlerquote je nach Klasse sehr unterschiedlich ist. Bei den Themen EU & Europa, Internationale Beziehungen & Armee sowie der Klasse «andere Probleme» liegt diese bei rund 30 %.

	Deutsch			Französisch		
Thema	Präzision	Sensitivität	F1-Mass	Präzision	Sensitivität	F1-Mass
Landwirtschaft	0.91	79.92%	0.85	94.64%	0.85	0.89
Gesundheitswesen	0.90	86.52%	0.88	95.68%	0.89	0.92
Erziehung & Kultur	0.87	80.49%	0.83	86.24%	0.86	0.86
Umwelt & Energie	0.84	81.69%	0.83	86.49%	0.88	0.87
Öffentliche Dienste / Infrastruktur	0.83	81.54%	0.82	90.73%	0.88	0.89
Wirtschaft	0.83	83%	0.83	90.84%	0.88	0.89
Immigration & Asyl	0.81	77.19%	0.79	85%	0.80	0.83
Finanzen & Steuern	0.81	77.30%	0.79	84.71%	0.86	0.85
Politisches System	0.81	69.66%	0.75	85.85%	0.82	0.84
Sozialversicherung & Sozialstaat	0.79	80.15%	0.80	86.71%	0.84	0.85
Geschlechterfragen & Diskriminierung	0.78	84.03%	0.81	82.57%	0.91	0.87
Recht & Ordnung	0.77	76.69%	0.77	95.68%	0.92	0.94
Internationale Beziehungen & Armee	0.75	72.81%	0.74	82.18%	0.77	0.79
Andere Probleme	0.75	74.22%	0.74	86.88%	0.80	0.83
EU / Europa	0.73	78.79%	0.76	77.18%	0.85	0.81
Arbeitsmarkt	0.71	77.00%	0.74	91.03%	0.85	0.88
Regionen & Nationaler Zusammenhalt	0.71	0.7158416	0.71	0.833713	0.78	0.81
Nicht klassifizierbar	0.49	0.6736232	0.57	0.7560976	0.82	0.79
Alle Artikel	0.78	0.777507	0.78	0.8675086	0.85	0.86

Tabelle 17: Evaluationstabelle der Ensemble-Methode für Zeitungsartikel

Mit der Ensemble-Methode ist es uns nun möglich, die Fehlerquote noch weiter zu reduzieren und eine Genauigkeit von 80 % und mehr für alle Klassen zu erreichen. Das führt auch dazu, dass die Präzision, die Sensitivität und das F1-Mass der Klassifikation

über die meisten Klassen hinweg sehr gut ausfällt. Das F1-Mass steht dabei für eine kombinierte Bewertung von Sensitivität und Präzision[1] der Methode.

In der Tabelle 17 präsentieren wir die Validationskennzahlen der Ensemble-Algorithmen für die Zeitungsartikel, wobei wir die Präzision, Sensitivität und das F1-Mass für alle Klassen (Themengebiete) zeigen. Tabelle 17zeigt hier sehr eindrücklich, dass mit der Verwendung des Ensemble-Modelles aus zwei verschiedenen Algorithmen und insgesamt drei Modellen einen durchschnittlichen F1-Wert von 0.78 für die deutschen Artikel und einen Wert von 0.87 für die französischen Artikel erreicht wird. Dies entspricht sehr guten Werten, vor allem für Französisch mit der kleineren Grösse des Datensatzes für das Training.

Geleichzeitig zeigt die Tabelle 17 auch, dass über alle Klassen hinweg gesehen die Ensemble-Methode sehr gut in beiden Sprachen funktioniert und nur gerade bei der Klasse der «nicht klassifizierbaren» Texte nicht sonderlich gut funktionieren. Diese Klasse umfasst allerdings recht wenige Texte. Darüber hinaus wird bei einer Gegenüberstellung der klassifizierten Häufigkeit der Themen mit der echten Häufigkeit (nach APS Trainingsdaten) ersichtlich, dass kein Thema durch die Klassifikation stark überrepräsentiert oder unterrepräsentiert wird im Vergleich zur Verteilung der Themen in den von der APS klassifizierten Trainingsdaten (max. -0.75-0.5 %).

Bei den Modellen zur Klassifikation der Beiträge aus den sozialen Medien entschieden wir uns für dieselbe Zusammensetzung wie bei den Zeitungsartikeln. Die Beweggründe sind dieselben. Denn auch bei den kurzen Texten aus den Sozialen Medien zeigt sich, dass die Deep-Learning Algorithmen eine jeweils eine Genauigkeit von mehr als 75 % auf, gefolgt von den besten Gradient-Boosting Modellen, die eine Genauigkeit von knapp 70 % erreichen. Wie schon bei den Zeitungsartikeln ist es auch hier mit der Ensemble-Methode möglich, die Fehlerquote noch weiter zu reduzieren und eine Genauigkeit von 84 % und mehr für alle Klassen zu erreichen. Das führt auch dazu, dass die Präzision, die Sensitivität und das F1-Mass der Klassifikation über die meisten Klassen hinweg sehr gut ausfällt.

Tabelle 18 zeigt dieselben Werte wie Tabelle 17, allerdings für die Ensemble-Methode der Beiträge aus den sozialen Medien.

[1] Eine Erklärung der Leistungsmasse finden Sie in Kelleher et al. (2015 S. 414 ff.)

	Deutsch			Französisch		
Thema	Präzision	Sensitivität	F1-Mass	Präzision	Sensitivität	F1-Mass
Nicht klassifiziert	97.39%	0.98	97.72%	0.94	0.96	0.95
Gesundheitswesen	95.26%	0.89	92.13%	0.54	0.84	0.65
Wahlen	91.39%	0.85	88.17%	0.90	0.69	0.78
Finanzen & Steuern	89.63%	0.88	88.95%	0.50	0.81	0.62
EU / Europa	89.19%	0.89	88.96%	0.64	0.85	0.73
Sozialversicherung & Sozialstaat	84.16%	0.89	86.55%	0.31	0.72	0.43
Umwelt & Energie	84%	0.88	85.56%	0.46	0.63	0.53
Immigration & Asyl	83.57%	0.91	87%	0.54	0.77	0.64
Geschlechterfragen & Diskriminierung	83.40%	0.91	86.86%	0.55	0.68	0.60
Abstimmung	68.03%	0.84	75.12%	0.47	0.79	0.59
Alle Tweets	86.56%	0.89	87.72%	0.58	0.77	0.65

Tabelle 18: Evaluationstabelle der Ensemble-Methode für Beiträge aus den sozialen Medien

Tabelle 18 zeigt, dass mit der Verwendung des Ensemble Modells aus zwei verschiedenen Algorithmen und drei Modellen ein durchschnittlicher F1-Wert von 0.88 für deutsche Socialmedia Beiträge erreicht wird, was ein sehr guter Wert ist und eine beträchtliche Steigerung gegenüber dem besten Deep-Learning Modell ist.

Geleichzeitig zeigt Tabelle 18 auch, dass über alle Klassen hinweg gesehen die Ensemble-Methode auf Deutsch sehr gut funktioniert. Bei den Beiträgen in Französisch sieht es weniger gut aus, da die Trainingsdatenmenge nicht genügend gross ist, um eine gute funktionierende Klassifikationsmethode zu trainieren. Der F1-Wert spiegelt dies mit einem Wert von nur 0.65 wider. Daher sind die Resultate der französischen Klassifikation mit Vorsicht zu geniessen. Deshalb zeigen wir im Bericht auch neben der Ensemble-Klassifikation die Klassifikation mit der manuellen Hashtag-Klassifikation, die zur Herstellung der Trainingsdaten benutzt wurde, womit zusätzlich überprüft werden kann, ob die Resultate des Ensemble-Algorithmus plausibel sind.

3.6 Named Entity Recognition (NER)

Named Entity Recognition ist ein maschinelles Annotationsverfahren, das nutzerspezifizierte Entitäten aus Texten extrahiert. Wir haben die NER primär dazu

verwendet, die Kandiderenden in den Zeitungsartikeln zu annotieren, um herauszufinden, welche Kandidierenden in welchen Artikeln erwähnt werden. Häufig werden für die NER komplizierte vortrainierte *machine learning-Algorithmen* (wie bspw. spaCy oder NTLK) verwendet, die Texte syntaktisch und lexikalisch nach verschiedenen Entiätstypen durchsuchen (z.B. geographische Entitäten oder Personen). In Fällen, in denen *ex ante* unklar ist, wonach in den Texten gesucht werden soll, ist eine Anwendung solcher Verfahren sinnvoll. Die Entitätserkennung mittels *machine learning-Algorithmen* führt jedoch teilweise zu Falschklassifikationen, die mit simpleren Methoden umgangen werden können. In unserem Fall existiert mit den Wahllisten ein Katalog von Personen, die wir in den Texten identifizieren wollen, sodass wir nicht umfassendere Annotationen bedürfen. Wir haben uns deshalb für ein konservatives dreistufiges Verfahren entschieden: In einem ersten Schritt werden die Namen aller Kandidierenden zu sogenannten *regular expressions* umcodiert. Dabei werden Zweitnamen und multiple Nachnamen zu optionalen Elementen, sodass sowohl nach Vollnamen und abgekürzten Alternativen gesucht wird. Im zweiten Schritt werden jeweils vier Sätze vor und nach der Erwähnung eines Namens extrahiert und in einem Korpus gesammelt. Danach werden diese Textstellen in einem dritten Schritt darauf überprüft, ob auch die Partei des darin erwähnten Kandidierenden vorkommt. Falls dies nicht der Fall ist, können wir nicht ausschliessen, dass es sich bei der Erwähnung auch um eine gleichnamige Person handeln könnte, und nicht um die Kandidatin. Dieser letzte Schritt dient der Vorbeugung von *false positives*, d.h. fälschlicherweise gematchten Textstellen.

3.7 Netzwerkanalyse

Für die Netzwerkanalyse verwenden wir eine Methode von Nocaj et al. (2015). Das Problem von Netzwerkanalysen in sozialen Netzwerken ist, dass die Abstände der vielen AkteurInnen in einem Netzwerk oft viel zu kleine durchschnittliche Entfernungen besitzen. Dies führt dabei oft zu *"Hairball"*-Netzwerken, oder auf Deutsch zu einem Wollknäuel. Dies kann zu den falschen Annahmen von einem hohem Vernetzungsgrad aller AkteurInnen miteinander führen. Dies insbesondere, wenn es mehrere Gruppen von stark vernetzen AkteurInnen gibt, die miteinander über wenige Verbindungen verfügen. Diese Struktur ist bei Netzwerkanalysen durchaus beabsichtigt, da es normalerweise das

eigentliche Ziel dieser Methoden ist, global einheitliche Kantenlängen zwischen den Knoten abzubilden.

Das Problem dabei ist, dass bei Diagrammen mit z.B. hoher Dichte oder in Gegenwart von Knoten mit hohem Vernetzungsgrad die Netzwerke dazu neigen, Knoten zusammenzuziehen. Dies führt dann wiederum zu Zuviel "Unordnung" im Netzwerk. Nocaj et al. (2015) schlagen gegen dieses Problem eine Methode vor, die speziell für eine Klasse von *Small-World-Diagrammen* entwickelt wurde, die typisch für soziale Online-Netzwerke sind. Das Verfahren basiert auf einem *übergreifenden Teilgraph*, der spärlich verbunden, aber noch verbunden ist und aus starken Bindungen besteht. Um diese Verbindungen zu identifizieren, nutzen sie ein Kriterium für die *strukturelle Einbettung,* was es ermöglicht Akteure mit hohem Vernetzungsgrad zu identifizieren. Diese werden daraufhin benutzt, um das Zentrum des Netzwerkes zu konstruieren, was zu einem übersichtlicheren Netzwerk führt. Eine Auswertung in empirischen und generierten Netzwerken von Nocaj, Ortmann, Brandes (2015) zeigt, dass ihr Ansatz die bisherigen Methoden verbessert.

4 Referenzen

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

Gilardi, F., Dermont, C., Kubli, M., Baumgartner, L., (2020), Der Wahlkampf 2019 in traditionellen und digitalen Medien, Zurich: IPZ, DigDemLab

Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006, May). A closer look at skip-gram modelling. In LREC (pp. 1222-1225).

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Nocaj, A., Ortmann, M., & Brandes, U. (2015). Untangling the hairballs of multi-centered, small-world online social media networks. Journal of Graph Algorithms and Applications: JGAA, 19(2), 595-618.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Proksch, Sven Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches." Legislative Studies Quarterly (September): 97–131.

Young, Lori, and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." Political Communication 29(2): 205–31.

5 Appendix

5.1 Liste der SMD Medien Quellen

Variable Name	Beschreibung
so	Abkürzung des Namesn der Zeitung
So_txt	Name der Zeitung
la	Publikationssparache des Artikels
n	Anzahl Artikel

so	so_txt	la	n
AGE	Agefi	fr	10014
APPZ	Appenzeller Zeitung	de	24422
ARC	Arcinfo	fr	11712
AVU	Anzeiger von Uster	de	869
AZM	Aargauer Zeitung / MLZ	de	23204
BAZ	Basler Zeitung	de	18935
BEOL	Berner Oberländer	de	18215
BEOL	Berner Oberländer	fr	1
BIT	Bieler Tagblatt	de	15235
BIT	Bieler Tagblatt	fr	16
BIZO	Bilanz online	de	937
BLI	Blick	de	11133
BODU	Bote der Urschweiz	de	19469
BU	Der Bund	de	14005
BUET	Bündner Tagblatt	de	13254
BZ	Berner Zeitung	de	16593
BZM	Basellandschaftliche Zeitung / MLZ	de	9977
CASO	Cash Online	de	42461
COOF	Coopération	fr	2075
COOI	Cooperazione	it	1851
COOP	Coopzeitung	de	2208
COOP	Coopzeitung	rm	1
FN	Freiburger Nachrichten	de	15035
FN	Freiburger Nachrichten	fr	6
FURT	Furttaler	de	1424
FUW	Finanz und Wirtschaft	de	3725
FUWO	Finanz und Wirtschaft	de	5295
FUWO	Finanz und Wirtschaft	en	95
GHI	GHI	fr	1136

GLAT	Glattaler	de	1201
GP	Glückspost	de	3499
HEU	24 heures	fr	15842
ILLE	L'Illustré	fr	1914
INFS	Infosperber	de	872
INFS	Infosperber	fr	2
JJ	Le Journal du Jura	fr	13441
LB	Der Landbote	de	16802
LBH	La Broye	fr	2756
LIB	La Liberté	fr	20260
LTZ	Limmattaler Zeitung / MLZ	de	12745
LUZ	Luzerner Zeitung	de	21829
MEWO	Medienwoche	de	125
MM	Migros-Magazin	de	2649
MME	Migros Magazine	fr	2003
NIW	Nidwaldner Zeitung	de	21342
NLZS	Zentralschweiz am Sonntag	de	2484
NNBE	Berner Zeitung	de	27346
NNBS	Basler Zeitung	de	19728
NNBS	Basler Zeitung	fr	1
NNBU	Der Bund	de	20193
NNHEU	24 heures	fr	28960
NNTA	Tages-Anzeiger	de	23550
NNTDG	Der Bund	fr	7
NNTDG	Tribune de Genève	fr	28170
NNTLM	Le Matin	fr	29086
NOU	Le Nouvelliste	fr	14589
NZZ	Neue Zürcher Zeitung	de	21068
NZZS	NZZ am Sonntag	de	4994
OAS	Ostschweiz am Sonntag	de	1870
OBW	Obwaldner Zeitung	de	21509
OLT	Oltner Tagblatt / MLZ	de	8553
ONA	Obersee Nachrichten	de	1325
RTS	rts.ch	fr	9901
RUEM	Rümlanger	de	1549
SBAU	Schweizer Bauer	de	6409
SBLI	Sonntagsblick	de	4156
SEBO	Seetaler Bote	de	2500
SF	Schweizer Familie	de	1226
SGT	St. Galler Tagblatt	de	25627
SHZ	Handelszeitung	de	2213
SHZO	Handelszeitung	de	4039

SI	Schweizer Illustrierte	de	1954
SOS	Südostschweiz	de	17176
SOS	Südostschweiz	it	7
SOS	Südostschweiz	rm	29
SOZM	Solothurner Zeitung / MLZ	de	15485
SRF	srf.ch	de	25616
SWII	swissinfo.ch	de	850
SWII	swissinfo.ch	en	851
SWII	swissinfo.ch	fr	853
SWII	swissinfo.ch	it	831
TA	Tages-Anzeiger	de	16986
TAGZ	Tagblatt der Stadt Zürich	de	2119
TAM	Das Magazin	de	560
TAS	SonntagsZeitung	de	3929
TBT	Toggenburger Tagblatt	de	23806
TDG	Tribune de Genève	fr	13588
TLMD	Le Matin Dimanche	fr	4495
TPS	Le Temps	fr	10742
TZ	Thurgauer Zeitung	de	29942
URZ	Urner Zeitung	de	21204
VOLK	Volketswiler	de	74
WASO	watson.ch	de	15068
WASO	watson.ch	en	1
WB	Walliser Bote	de	18929
WEOB	Werdenberger & Obertoggenburger	de	17336
WEW	Die Weltwoche	de	2437
WILB	Willisauer Bote	de	5942
WOZ	Die Wochenzeitung	de	1466
ZHOL	Zürcher Oberländer	de	16831
ZHUL	Zürcher Unterländer	de	15148
ZOF	Zofinger Tagblatt / MLZ	de	14232
ZPLU	Zentralplus	de	7836
ZSZ	Zürichsee-Zeitung	de	17121
ZUGZ	Zuger Zeitung	de	23035
ZWA	20 minuten	de	18662
ZWAI	20 minuti	it	8660
ZWAO	20 minuten	de	20453
ZWAS	20 minutes	fr	15161

5.2 Liste der Themen der Zeitungsartikel aus der SMD und des CdT

Thema	Themen Beschreibung der Artikel
Nicht klassifiziert / Nicht politisch	Alle Texte, die keiner Klasse zugewiesen wurden, da entweder nicht auf Deutsch oder Französisch oder keinen politischen Inhalt aufweisen.
EU / Europa	Alle politischen Texte zum Verhältnis der Schweiz mit der EU und Europa
Umwelt & Energie	Alle politischen Texte die sich mit dem Thema Umwelt und Energie befassen
Erziehung & Kultur	Alle Texte die sich um Fragen zur Bildungspolitik, Forschungspolitik, Gesellschaft und Kunst und Kultur Förderung drehen
Wirtschaft	Artikel die sich mit der Wirtschaftspolitik auseinander setzen
International Beziehungen & Armee	Dieses Thema beinhaltet alles was mit internationaler Politik und der Schweiz zu tun hat, sowie die Sicherheitspolitik
Einwanderung & Asyl	Dies umfasst alle politischen Texte, die sich mit der Einwanderungs- und der Asylpolitik auseinander setzen
Landwirtschaft	Dieses Thema beinhaltet Texte zur Agrarpolitik der Schweiz. Dies beinhaltet auch das Forstwesen, die Fischerei und die Ernährungspolitik
Sozialversicherung / Sozialstaat	Beinhaltet alle Artikel, die sich mit dem Sozialstaat auseinandersetze, wie zum Beispiel die Altersvorsorge, Sozialhilfe und dem Rentenalter
Geschlechterfragen & Diskriminierung	Dieses Thema umfasst alle Texte zur Gleichstellung zwischen Mann und Frau, sowie Texte, die sich mit Diskriminierung befassen
Arbeitsmarkt	Dieses Thema umfasst alle Debatten zum Arbeitsmarkt.
Recht & Ordnung	Dieses Thema umfasst neben Diskussionen zum Rechtssystem auch sehr viele Texte, die von wichtigen Prozessen in der Schweiz handeln und auch Texten, die sich mit dem Sicherheitsapparat der Polizei auseinandersetzen
Finanzen & Steuern	Alle Artikel, die mit der Finanzpolitik zu tun haben, fallen in dieses Thema
Gesundheitswesen	Das Gesundheitswesen umfasst Krankenversicherungsdiskussionen, sowie auch Medizinische Themen.
Öffentliche Dienste & Infrastruktur	Dieses Thema ist ein Sammelbecken für fast alles. Es Umfasst alle Texte, die mit öffentlichen Bauvorhaben zu tun haben, sowie alles was sich um den Öffentlichen Verkehr, den Individualverkehr und Lastenverkehr dreht. Dies Inkludiert auch Lärmpolitische Diskussionen
Regionaler & nationaler Zusammenhalt	Dieses Thema beinhaltet alle Texte die sich mit der Schweiz als Einheit auseinandersetzen.
Politisches System	Dieses Thema beinhaltet alle Texte die sich mit Wahlen, Abstimmungen, Parteien, Behörden auseinandersetzen
Andere Probleme	Alles was politisch ist und zu keinem die obigen Themen passt

5.3 Liste der Themen zu den Beiträgen aus den sozialen Medien

Thema	Themen Beschreibung der Artikel
Nicht klassifiziert	Alle Texte, die keiner Klasse zugewiesen wurden, da entweder nicht auf deutsch oder Französisch oder keinen politischen Inhalt aufweisen
Gesundheitswesen	Das Gesundheitswesen umfasst Krankenversicherungsdiskussionen, sowie auch Medizinische Themen
Wahlen	Alle Texte mit dem Inhalt Wahlen
Finanzen & Steuern	Alle Artikel, die mit der Finanzpolitik zu tun haben, fallen in dieses Thema
EU / Europa	Alle politischen Texte zum Verhältnis der Schweiz mit der EU und Europa
Sozialversicherung & Sozialstaat	Beinhaltet alle Artikel, die sich mit dem Sozialstaat auseinandersetze, wie zum Beispiel die Altersvorsorge, Sozialhilfe und dem Rentenalter
Umwelt & Energie	Alle politischen Texte die sich mit dem Thema Umwelt und Energie befassen
Immigration & Asyl	Dies umfasst alle politischen Texte, die sich mit der Einwanderungs- und der Asylpolitik auseinander setzen
Geschlechterfragen & Diskriminierung	Dieses Thema umfasst alle Texte zur Gleichstellung zwischen Mann und Frau, sowie Texte, die sich mit Diskriminierung befassen
Abstimmung	Dieses Thema umfasst alle Texte, die zu den Abstimmungsdebatten beitragen.